Furthermore, modifications to nucleotidic units include rearranging, appending, substituting for or otherwise altering functional groups on the purine or pyrimidine base which form hydrogen bonds to a respective complementary pyrimidine or purine. The resultant modified nucleotidic unit optionally may form a base pair with other such modified nucleotidic units but not with A, T, C, G or U. Abasic sites may be incorporated which do not prevent the function of the polynucleotide. Some or all of the residues in the polynucleotide can optionally be modified in one or more ways.

Standard A-T and G-C base pairs form under conditions which allow the formation of hydrogen bonds between the N3-H and C4-oxy of thymidine and the Nl and C6-NH2, respectively, of adenosine and between the C2-oxy, N3 and C4-NH2, of cytidine and the C2-NH2, N'-H and C6-oxy, respectively, of guanosine. Thus, for example, guanosine (2-amino-6-oxy-9-β-D-ribofuranosyl-purine) may be modified to form isoguanosine (2-oxy-6-amino-9-β-D-ribofuranosyl-purine). Such modification results in a nucleoside base which will no longer effectively form a standard base pair with cytosine. However, modification of cytosine (1-β-D-ribofuranosyl-2-oxy-4-amino-pyrimidine) to form isocytosine (l-β-D-ribofuranosyl-2-amino-4-oxy-pyrimidine) results in a modified nucleotide which will not effectively base pair with guanosine but will form a base pair with isoguanosine (U.S. Pat. No. 5,681,702 to Collins et al.). Isocytosine is available from Sigma Chemical Co. (St. Louis, MO); isocytidine may be prepared by the method described by Switzer et al. (1993) Biochemistry 32:10489-10496 and references cited therein; 2'-deoxy-5-methyl-isocytidine may be prepared by the method of Tor et al. (1993) J. Am. Chem. Soc. 115:4461-4467 and references cited therein; and isoguanine nucleotides may be prepared using the method described by Switzer et al. (1993), supra, and Mantsch et al. (1993) Biochem. 14:5593-5601, or by the method described in U.S. Patent No. 5,780,610 to Collins et al. Other nonnatural base pairs may be synthesized by the method described in Piccirilli et al. (1990) Nature 343:33-37 for the synthesis of 2,6-diaminopyrimidine and its complement (1-methylpyrazolo-[4,3]pyrimidine-5,7-(4H,6H)-dione. Other such modified nucleotidic units which form unique base pairs are known, such as those described in Leach et al. (1992) J. Am. Chem. Soc. 114:3675-3683 and Switzer et al., supra.

The phrase "DNA sequence" refers to a contiguous nucleic acid sequence. The sequence can be either single stranded or double stranded, DNA or RNA, but double stranded

DNA sequences are preferable. The sequence can be an oligonucleotide of 6 to 20 nucleotides in length to a full length genomic sequence of thousands of base pairs.

A "library of DNA sequences" refers to a plurality of DNA sequences. The number of "members of the library" is not critical; it can range from less than ten to greater than 10^6. Typically in a library of DNA sequences, the library contains many different DNA sequences, all derived from the same parent DNA sequence but containing mutations in the sequence. The phrase "creating a library of DNA sequences" refers to the physical generation of a library of DNA sequences. Techniques used to physically generate a library are well know in the art and are referenced below. Typically, a "phage library" is created. "Phage libraries" comprise a DNA library incorporated into bacteriophage. The library is constructed such that the proteins encoded by the DNA library are expressed on the surface of the phage and thus on the surface of infected bacteria. The bacteria which contains the library is then "screened" for the presence of proteins with desired functionality. A "second library" is a library of DNA sequences based on the results found in the first library of DNA sequences. For example, if a beneficial mutation is found in the screening of a library, the mutation may be incorporated into the protein upon which the second library is based.

The term "IRL" refers to an information-rich library such as produced by a method of the invention.

The term "protein" refers to contiguous "amino acids" or amino acid "residues." Typically, proteins have a function. However, for purposes of this invention, proteins also encompasses polypeptides and smaller contiguous amino acid sequences that do not have a functional activity. The functional proteins of this invention include, but are not limited to, esterases, dehydrogenases, hydrolases, oxidoreductases, transferases, lyases, and ligases. Useful general classes of enzymes include, but are not limited to, proteases, cellulases, lipases, hemicellulases, laccases, amylases, glucoamylases, esterases, lactases, polygalacturonases, galactosidases, ligninases, oxidases, peroxidases, glucose isomerases and any enzyme for which closely related and less stable homologs exist. In addition to enzymes, the encoded proteins which can be used in this invention include, but are not limited to, transcription factors, antibodies, receptors, growth factors (any of the PDGFs, EGFs, FGFs, SCF, HGF, TGFs, TNFs,

insulin, IGFs, LIFs, oncostatins, and CSFs), immunomodulators, peptide hormones, cytokines, integrins, interleukins, adhesion molecules, thrombomodulatory molecules, protease inhibitors, angiostatins, defensins, cluster of differentiation antigens, interferons, chemokines, antigens including those from infectious viruses and organisms, oncogene products, thrombopoietin, erythropoietin, tissue plasminogen activator, and any other biologically active protein which is desired for use in a clinical, diagnostic or veterinary setting. All of these proteins are well defined in the literature and are so defined herein. Also included are deletion mutants of such proteins, individual domains of such proteins, fusion proteins made from such proteins, and mixtures of such proteins; particularly useful are those which have increased half-lives and/or increased activity.

"Polypeptide" and "protein" are used interchangeably herein and include a molecular chain of amino acids linked through peptide bonds. The terms do not refer to a specific length of the product. Thus, "peptides," "oligopeptides," and "proteins" are included within the definition of polypeptide. The terms include polypeptides contain co- and/or post-translational modifications of the polypeptide, for example, glycosylations, acetylations, phosphorylations, and sulphations. In addition, protein fragments, analogs (including amino acids not encoded by the genetic code, e.g. homocysteine, ornithine, D-amino acids, and creatine), natural or artificial mutants or variants or combinations thereof, fusion proteins, derivatized residues (e.g. alkylation of amine groups, acetylations or esterifications of carboxyl groups) and the like are included within the meaning of polypeptide.

"Amino acids" or "amino acid residues" may be referred to herein by either their commonly known three letter symbols or by the one-letter symbols recommended by the IUPAC-IUB Biochemical Nomenclature Commission. Nucleotides, likewise, may be referred to by their commonly accepted single-letter codes.

"Variants of a protein" are those proteins that are related to one another by a common amino acid sequence or "parental protein" but contain minor variations in amino acid sequence from each other. These changes can be conservative substitutions, non-conservative substitutions, deletions, insertions or substitutions with non-naturally occurring amino acids (mimetics). The phrase "optimizing a protein" refers to the process of changing a protein to protein variants so that the desired functionality is improved. One of skill will realize that